CyberPeace editor@cyberpeace.net

THE CYBERPEACE JOURNAL

Supported by United Service Institution of India

Academic Partner Shah & Anchor Kutchhi Engineering College

Visual Speech: A Deep Learning Framework for Robust Lipreading

Jayan Shah
Cyber Security
Shah & Anchor Kutchhi Engineering College
Mumbai, India
https://orcid.org/0009-0000-9677-9175

Pratham Shah
Cyber Security
Shah & Anchor Kutchhi Engineering College
Mumbai, India
https://orcid.org/0009-0006-0935-6865

Smit Borkhetaria
Cyber Security
Shah & Anchor Kutchhi Engineering College
Mumbai, India
smit.17436@sakec.ac.in

Prajakta Pote
Cyber Security
Shah & Anchor Kutchhi Engineering College
Mumbai, India
prajakta.pote@sakec.ac.in

***Abstract*—Lipreading, the capacity to decipher spoken words from lip movements, has a wide range of uses, including silent communication, improving multimedia speech recognition, and helping the hard of hearing. Nonetheless, lipreading continues to be a difficult undertaking because of the intricacy and unpredictability of lip motions in addition to the scarcity of big training datasetsIn this work, we provide a unique deep learning model for visual speech recognition that achieves state-of-the-art performance on popular lipreading benchmarks.**

Keywords—Convolutional Neural Network, Lipreading, LSTM

I. INTRODUCTION

The process of comprehending speech by visually analyzing lip, face, and tongue movements is called lip reading, or visual speech recognition. Its significance has grown in a number of applications, including as assistive technology for people with hearing loss, speech recognition in loud situations, and human-computer interaction. However, due to the intricacy of lip motions, speaker unpredictability, and the existence of occlusions brought on by accessories or facial features, reliable lip reading is still a difficult endeavor. Researchers have investigated a variety of strategies over time to address the lipreading issue, from conventional machine learning methods to deep learning models.[1] Early research mapped visual features based on lip patches to phonemes or words using neural networks and classification algorithms, including Hidden Markov Models (HMMs) and Support Vector Machines (SVMs).

These techniques could not, however, fully capture the complex patterns of lip movements or the contextual data required for precise identification.The subject of lip reading has advanced significantly with the introduction of deep learning. Deep neural networks outperform conventional feature-based techniques for acquiring representations directly from raw footage, as evidenced by groundbreakingly studies by Petridis et al. and Chung and Zisserman. To further enhance lip reading performance, later research has looked into a variety of designs, including LSTM networks (Long Short-Term Memory), Convolutional Neural Networks [2] and multimodal fusion approaches. Even with these improvements, a number of issues still need to be resolved, such as managing speaker variability, addressing co-articulation effects, and adding context for more reliable identification. Furthermore, the creation and assessment of increasingly advanced models have been hampered by the absence of extensive, varied lipreading datasets.

In order to tackle some of these issues, this study offers "LipSync," a revolutionary deep learning method for visual speech detection. Our suggested approach aims to push the limits of reading lips accuracy and robustness by fusing cutting-edge neural network designs with creative methods for capturing changes in time and adding contextual clues. In addition, we investigate methods for multi-task learning, transfer learning, and data augmentation to strengthen generalization and counteract overfitting. Our comprehensive tests on both our proprietary dataset gathered in real-world scenarios and publicly accessible lipreading datasets show our approach's superiority over current techniques.

II. OBJECTIVE

Developing an end-to-end machine learning algorithm for automated lip-reading that can precisely parse text from a speaker's lip movements in a video is the main

goal of the LipSync project. The main objectives of the research are:

1. To create an end-to-end trainable model that can directly map variable-length video sequences of lip movements to the corresponding text, without the need for separate feature extraction and prediction stages.
2. To design a model architecture that effectively captures the temporal dynamics and contextual information present in lip movements, which is essential for accurate lipreading, especially for longer words and continuous speech.
3. To develop a reliable and effective lipreading model by utilizing cutting-edge deep learning methods such spatiotemporal convolutions in general recurrent artificial neural networks, and the connectionist temporal classification loss. To enable the model to perform sentence-level sequence prediction, going beyond word classification tasks, and thereby enabling lipreading for continuous speech and natural language processing applications.
4. To advance the field of automatic lipreading by developing a model that outperforms existing approaches and sets a new benchmark for accuracy and performance.

III. OBJECTIVES

1. To overcome the limitations of traditional lipreading approaches, which relied on separate stages for feature extraction and prediction, by proposing an end-to-end trainable model that can map different-length video sequences to text.
2. To address the challenge of capturing temporal context and dynamics in lip movements, which is crucial for accurate lipreading, especially for longer words. The LipSync model aims to achieve this by incorporating spatiotemporal convolutions and recurrent networks in its architecture.
3. To leverage the connectionist temporal classification loss function, which provides the algorithm to be trained completely end-to-end, without the need for pre-processing data or post-processing steps.
4. To enhance current deep lip-reading designs, which often carry out word categorization tasks but do not deal with sequence prediction at the phrase level. With possible applications in areas like multimedia analysis, assistive technology for the deaf, and human-computer interaction, the LipSync project seeks to improve advances in automatic lipreading by attaining these goals.

IV. PROPOSED ARCHITECTURE

A. Architecture

The connected 3D system architecture necessitates training two distinct networks with various weight configurations. The temporal and spatial information obtained from lip movements are combined and mixed for the sight networks in order to take use of the temporal correlation. The stacking audio frames form the audio network's temporal dimension, while the energy properties that are collected are considered to be its spatial dimension. In our proposed 3D CNN architecture, the neural processes for both audio-visual streams are performed on sequential temporal frames. All layers with full connectivity have used dropout(ρ) up until the last layer. As proposed, PReLU activation comes after every layer except the final one, generalising to ReLU.

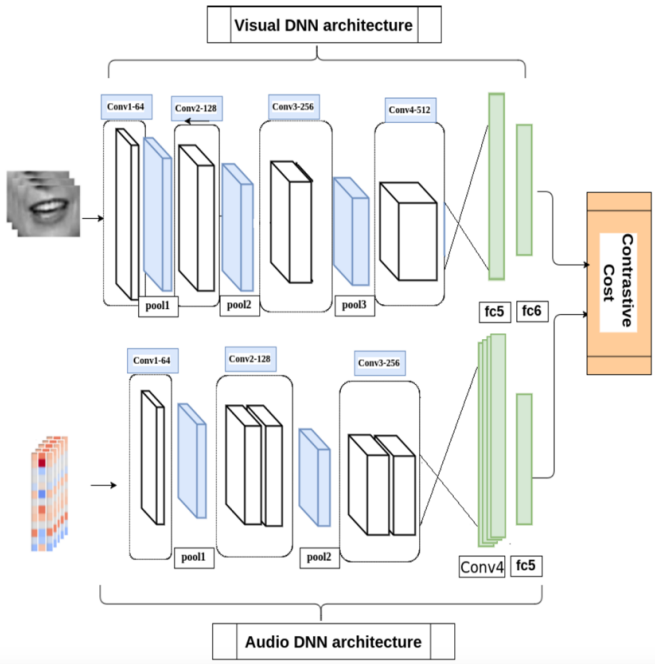

Fig 1. Architecture of Audio-Video Model

The network architecture used for streaming video training is shown in figure 1 above. The temporal kernel dimension is denoted by T, and the width and height kernel sizes are indicated by H and W, respectively, in Table 1's dimensional representation of the 3D kernels. In conventional CNN systems, the kernel depth is determined by the input channel size or the number of feature crossings in the previous layer. The kernel dimension representation does not employ any kernel depth values for simplicity's sake. One important component of the optical system is its pooling mechanism. Since we are using 3D convolutional layers, we also need 3D pooling layers. When using 1x3x3 kernels for spatial data pooling, the number of pools step is adjusted to two in order to improve robustness against the impact of moving lips and maintain lip motion features in the region of the pooling kernel. A high degree temporal as well as spatial information are associated through its fusion using 3D convolutional techniques. The zero-padding technique is not utilised in the visual architecture. We only apply pooling operations on the frequency dimension (domain) in our design in order to maintain the temporal data under the time frames. Moreover, our proposed design has a high degree of compression and only requires 64 units of output.

To extract spatiotemporal features, the first layer combines a 3D kernels with a 3D convolutional technique. With the sole exception of the initial layer, we basically facing 2D dimensionality because the depth dimension for

sound map features in higher-level layers is M × N × 1. Because of this, the audio network features frequent 2D convolutional processes that simultaneously gather temporal and spatial input by utilising their 2D kernels. The deliberate representation of the basic elements of spacetime as T × H × 1 highlights the relationship between convolutional processes in two and three dimensions. Zero-padding is not employed in the audio architecture because it introduces additional fictitious zero-energy coefficients that have no bearing on local feature extraction. Another crucial element is the application of non-square kernels.

The kernel widths decrease in the following order: minimal characteristics to high-level features. This method yields correlated a high degree features, or the features retrieved from the CNN, as well as additional temporal information in the lower categories that are linked to speech aspects. CNNs, or convolutional neural networks, have become much more efficient at computer vision applications that use images as input. Convolutions that are stacked and act spatially throughout an image make up CNNs.

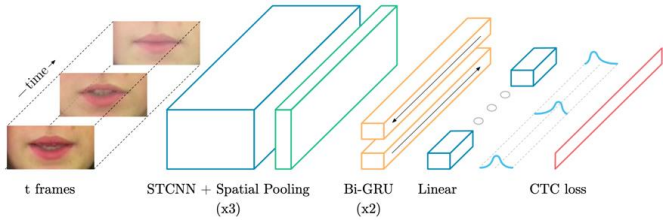

Fig 2. Architecture of Spatiotemporal Convolutions and Gated Recurrent Unit

As shown in figure 2 above, three layers of STCNN are utilized to process an order of T frames, with a spatial maximum pooling layer coming after each layer. Two Bi-GRUs process the retrieved features, and a layer of linearity and a soft max process every phase of the GRU output. CTC is utilized to train this end to-end model. Two layers fits the audio video model which are:

$$[\mathrm{conv}(\mathbf{x},\mathbf{w})]_{c'ij} = \sum_{c=1}^{C}\sum_{i'=1}^{k_w}\sum_{j'=1}^{k_h} w_{c'ci'j'}x_{c,i+i',j+j'}$$

Fig 3. Equation of STCNNs layer used

$Xc(ij) = 0\ for\ i,j$ out of limits where we define the weights $w \in R\ C\ 0 \times C \times kw \times kh$ for input x. By convolving across time and spatial dimensions, spatial convolution neural networks (STCNNs) are able to analyse video input.

$$[\mathbf{u}_t, \mathbf{r}_t]^T = \mathrm{sigm}(\mathbf{W}_z\mathbf{z}_t + \mathbf{W}_h\mathbf{h}_{t-1} + \mathbf{b}_g)$$
$$\tilde{\mathbf{h}}_t = \tanh(\mathbf{U}_z\mathbf{z}_t + \mathbf{U}_h(\mathbf{r}_t \odot \mathbf{h}_{t-1}) + \mathbf{b}_h)$$
$$\mathbf{h}_t = (\mathbf{1} - \mathbf{u}_t) \odot \mathbf{h}_{t-1} + \mathbf{u}_t \odot \tilde{\mathbf{h}}_t$$

Fig 4. Equation of Gated Recurrent Unit

where $z : \{z1, \ldots, zT\}$ is input to the RNN, denotes element-wise multiplication, and $sigm(r) = 1/(1 + exp(-r))$.

V. PROPOSED ARCHITECTURE

A. The Problems and the Approach

The primary challenge, and the aim of this effort, is to determine the relationship among both visual and audio streams. In order to assess the correlation of visual-audio signals utilizing the different modal traits that have been learned, we suggested utilizing a linked 3D CNN (Convolutional Neural Network) architecture that is capable of mapping both modes into a representation space.

B. Dataset Used

The WVU Audio-Visual Dataset (AVD) and Lip Reading in the Wild (LRW) are the datasets that have been employed in our investigations. The LRW dataset includes up to thousand separate speaker's utterances of five hundred distinct words. Every video last for 1.16 seconds, and the word appears midway through. The video and audio data that were gathered between 2022 and 2013 make up the AVD dataset. Both planned and unstructured voice samples are included in the audio and video material. The participant recited a passage from the scripted samples. In contrast to giving a straightforward "yes" or "no" response, the participant in the unscripted samples responded to interview questions that elicited conversational answers.

C. Processing

Fig. 1 below displays the pipeline used to process both datasets. There are two visual and auditory portions to the pipeline. The videos in the visual part are post-segmented to have a constant rate of thirty frame/s. Next, using the dlib library, face tracking and mouth area extraction are done on the videos. The input feature cube is created by concatenating all of the mouth areas after they have been shrunk to the same size. There are no audio files in the dataset. The FFmpeg framework is used in the audio division to extract audio sections from videos. After that, audio files will have their speech characteristics extracted. SpeechPy is the library which is being utilised for the task of extracting speech features.

D. Data Representation

A pair of audio and video streams are used by the two non-identical ConvNets in the suggested design. Two characteristics that indicate speech and lip movement that were taken from a 0.3/second visual clip are used as the network input. Finding out if an audio stream matches a lip movement clip for the intended stream time is the primary task. The brief duration of the video clip (0.3–0.5 seconds) used to assess the method adds to the task's complexity. Because only a little quantity of recorded video or audio may be available for certain biometrics or forensics operations to distinguish between distinct modalities, this setting is similar to real-world situations. The temporal elements of audio and video must line up over the span of time they cover. The next two parts address this correspondence.

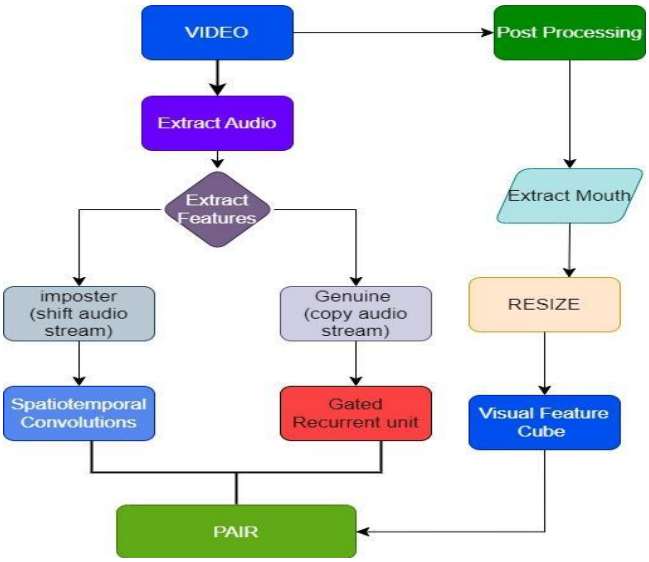

Fig 5. Flowchart of LipSync Model

A. Speech

CNNs are distinguished primarily by their localization, which refers to the application of the convolution process to particular local regions inside an image. There should be some sort of correlation between the neighbor features as a visual interpretation of this location property. When using a CNN design, the input voice feature maps are interpreted as pictures, hence the features on each axis must have a local correlation in terms of both time and frequency. The speech feature representation that can be employed is the MFCCs, obtained from the description of the sound stream. This is because the order of the filter streams is altered during the last motion (DCT)2 for creating MFCCs, which aims to remove relations between energy coefficients. This disturbs the locality attribute. The method used is based on the log-energies, or what we refer to as MFECs, that are directly obtained from the filter-bank energies.3. Similar to MFCCs, MFECs are extracted without the need for a DCT process. The temporal characteristics on the time axis are 20 ms non-overlapping frames that are used to generate spectrum features with a local characteristic.

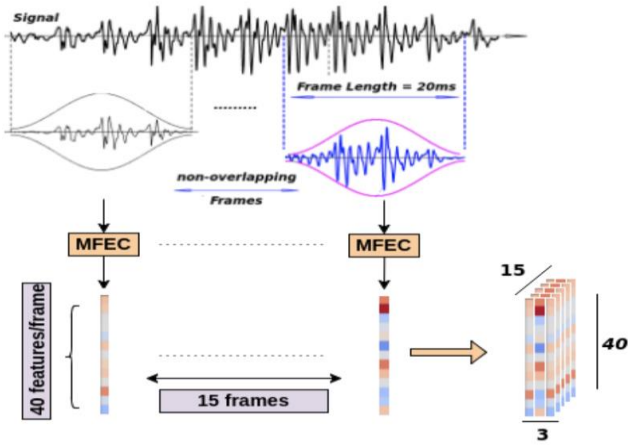

Fig 6. Stacking frames made from the incoming signal samples to generate speech

B. Video

Every video clip used in this effort has a 30 frames per sec frame rate. Consequently, nine consecutive image frames make up the 0.3-second visual stream. A cubic of 9 × 60 × 100 dimensions is sent into the visual stream of the network, where 9 is the number of temporal information-containing frames. Each channel has a grayscale image of the mouth region that measures 60 by 100 pixels. Figure 3 below provides an example of a mouth area representation. The relatively narrow lip cropping zone was chosen on purpose due to practical reasons, as high-resolution pictures are uncommon to be accessible in real-world situations. In addition, unlike traditional CNN experimental setups, we did not used the tests to pictures with static square aspect frames.

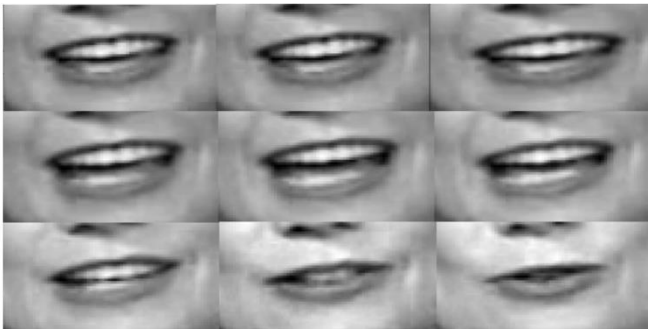

Fig 7. The sequence of mouth areas in a 0.3-second video stream

C. Data Augmentation

- *Pre-Processing*

34 subjects, each repeating 1000 sentences, make up the GRID corpus. 32746 of the movies are feasible out of all of them; nevertheless, the videos of speaker 21 are absent, and some of the other videos have been damaged or empty. For evaluation (3971 videos), we employ an allocation (undetected speakers; not utilised in prior research), excluding the data from two female presenters (20 and 22) and two male presenters (1 and 2). The remainder is earmarked for instructional purposes (28775 videos). In contrast, for the split (merged) speakers, we additionally use a sentence-level variation in which the 255 randomly selected phrases of each speaker are assessed. For training, all of the participant's leftover data is combined.

- *Augmentation*

To minimise overfitting, we make small changes to the dataset. First, traditional vertically mirrored pictures are used for training. Secondly, we add video clips of certain phrases as extra training instances to the phrase-level training information as the dataset gives word beginning and end timings for every sentence video. The degradation rate under these conditions is 0.925. Third, a per-frame rate of 0.05 is used to delete and duplicate frames in order to increase robustness against different motion speeds. All of the suggested models and starting points used the same augmentation techniques.

II. RESULTS

Evaluation on LRW Dataset: You give 500 words (subjects) for multimedia matching to the Lip Read in the Wild dataset. The first 400 words are used to create the training set, while the last 100 words are utilised to create the test set, so as to make both sets of tests mutually exclusive. Just 50 syllables from each sentence are selected for data output for each train/test set. Authentic and fake data pairings are created in the first training data compilation. Since not all of the generated data is used for training, this is referred to as initial training data. Figure 8 below summarises the feature testing and training procedures.

set	# subjects(words)	# word utterances	# pairs
train	400	50	280k
test	100	50	70k

Fig 8. Evaluation on LRW Dataset

Real pairs (audio/video) are produced by comparing the relevant sound feature cubic to the 9-convey visual feature cube. The voice feature map of a movie moves along its time axis to generate impostor pairs. The shifting is completely random and can go on for up to 0.5 seconds. The pair generation method is depicted in the following Fig. 9. Several studies have been carried out to examine the impact of design, feature selection, and pair selection methodology. In all of the studies that were conducted, we created imposter pairs using a 0.5-next shift in order to generate test data. In addition to MFEC features, the investigations included one and second order derivatives, and 64 was selected as the resultant root feature space dimension.

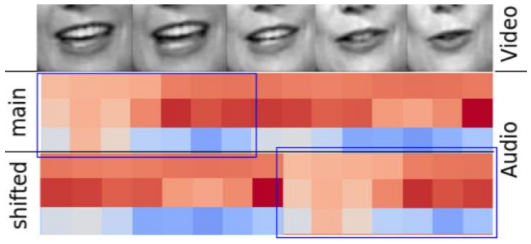

Fig 9. Audio and video feature maps for pair generation

In this instance, we present the outcomes regardless of the online partnership selection. Pair selection in the World Wide Web sped up the development of the training loss and improved accuracy. It also reached the maximum test accuracy faster. For each setup, the EER was calculated at several training epochs. Figure 6 displays the average results of five training cycles as well as the rate of diverging and performance gain when the EER is used to evaluate the test results for the full number of learning epochs. Fig. 10 below shows the impact on internet pair choice on dependability using the default configuration.

It is clear that the experimental condition with the least amount of temporal shift is the most difficult. The AVD dataset's ROC curve is displayed in Figure 11 below. Fine-tuning is expected because it has enhanced the similarity between the real and false pairs, which has a negative relationship to the time-shifted data used to create the fake pairs.

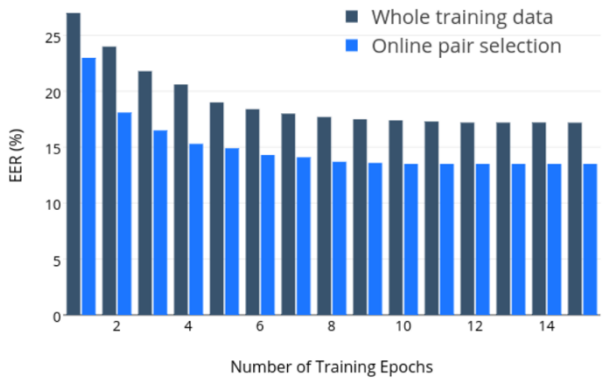

Fig 10. The effect of the proposed adaptive online pair selection

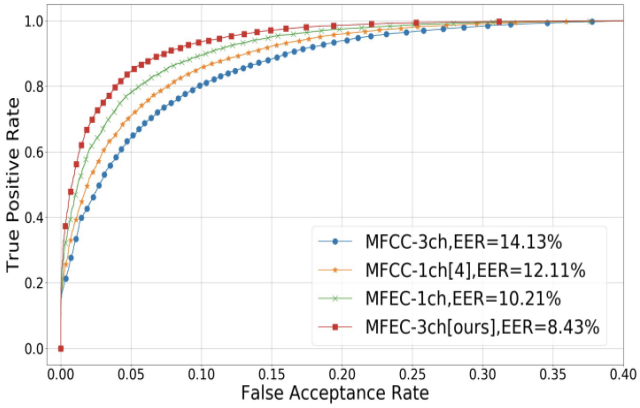

Fig 11. The ROC curve representation for fine-tuning on AVD dataset.

The training approach consists of ending work with the AVD dataset and fine-tuning the settings of the previously learned network after evaluating the Private Limited, No Outside Data configuration. To achieve optimisation, a decay-free rate of learning of 10−6 was employed together with 15 training epochs of training data.

Notably, using temporal variations for MFCC features led to a performance decrease, as Fig. 12 below illustrates. This can be related to the computation of the global derivative feature when using non-local MFCC features. The trials carried out on the AVD dataset demonstrate an improvement on the Equal Error Rates (EER) that exceeds 29% when comparing the proposed method with the state-of-art method.

III. FUTURE EXPLANATION

While the proposed LipSync model and coupled 3D convolutional architecture have shown promising results for automatic lipreading and audio-visual recognition, there are several avenues for further research and improvement.

A. *Multimodal Fusion:* Although this work explored the fusion of audio and visual modalities, incorporating additional modalities, such as facial expressions, head movements, or contextual information, could potentially enhance the model's performance and robustness. Investigating advanced fusion techniques and architectures that can effectively combine multiple modalities is an intriguing direction for future work.

B. *Continuous Speech Recognition:* The current LipSync model is designed for word-level lipreading tasks. Extending the model to handle continuous speech recognition, where it can generate transcripts for complete sentences or conversations, would significantly expand its practical applications. This extension would require addressing challenges such as handling co-articulation effects and incorporating language models.

C. *Transfer Learning and Domain Adaptation:* Variations in speaker quality, illumination, and background noise can all have an impact on the model's performance. Investigating domain adaption tactics and transfer learning 30 approaches may enhance the model's generalisation skills and permit reliable lipreading in a variety of real-world contexts.

D. *Attention Mechanisms:* Incorporating attention mechanisms into the LipSync model could allow it to focus on the most relevant spatial and temporal regions of the input video, potentially improving its accuracy and interpretability. Attention mechanisms could also aid in integrating contextual information and handling complex speech patterns.

IV. CONCLUSION

In this study, we present LipSync, a complete deep learning framework for automatic reading lips that directly translates variable-length video frame sequences to text. A recurrent network, spatiotemporal convolutions, and both connectionist chronological classification loss are used by LipSync to efficiently capture the dynamics and temporal context of lip movements—two essential components of accurate lipreading, particularly for lengthier words. Furthermore, using 3D convolutions and pool operations, we introduced a novel linked 3D convolutional design for multimedia stream fusion that includes temporal convolutional fusion. When compared to other approaches, this architecture dramatically lowers the number of variables while facilitating the effective combination of auditory and visual modalities.

Our proposed approaches were proven to be superior through extensive experiments conducted on multiple datasets. In word classification tasks, LipSync outperformed conventional and current deep lipreading models, while in audio-visual matching tasks, the combined 3D convolutional architecture achieved state-of-the-art performance. Our research emphasises the significance of multi-modal fusion and temporal modelling for precise audio-visual recognition and lipreading. The suggested techniques open up new possibilities for development in these difficult fields, with possible uses in multimedia analysis, assistive technology, and human-computer interaction.

ACKNOWLEDGMENT

We take this opportunity to express our sincere thanks to our Guide, Ms. Prajakta Pote a Faculty in the Department of Cyber Security in Shah and Anchor Kutchhi Engineering College for guiding us and suggesting regarding the line of work for our project “LipSync”. We would like to express our gratitude towards their constant encouragement, support and guidance throughout the progress.

Also, we would like to thank our Principal – Dr. Bhavesh Patel and Dr. Nilakshi Jain, Head of Cyber Security Department, for their help, support & guidance for this project. We are also thankful to all Faculty members of our department for their help and guidance during completion of our project.

REFERENCES

[1] Assael, Y.M., Shillingford, B., Whiteson, S. and de Freitas : “Deep Learning-Based Automated LipReading: A Survey” (2016)
[2] Assael, Y.M., Shillingford, B., Whiteson, S. and de Freitas : “LipNet: End-to-End Sentence-level Lipreading” (2020)
[3] Hu, D., Li, X. and Chuah, M.C : “3D Convolutional Neural Networks for Cross Audio-Visual Matching Recognition” (2021)
[4] Shillingford, B., Assael, Y., Hoffman, M.W., Paine, T., Hughes, C., Prabhat, U., Liao, H., Sak, H., Rao, K., Bennett, L. and Mulville : “Deep Lip Reading: a comparison of models and an online application” (2022)
[5] Shillingford, B., Assael, Y., Hoffman, M.W., Paine, T., Hughes, C., Prabhat, U., Liao, H., Sak, H., Rao, K., Bennett, L. and Mulville, M. : “Lip Reading Word Classification” (2022)
[6] Merkulova, A., Khomitsevich, O. and Vasilyeva : “Training Strategies for Improved Lip-Reading” (2022)
[7] Dixit, A., Rasiwasia, N. and Virmani, J : “Lip Reading Sentences Using Deep Learning With Only Visual Cues” (2021)
[8] Virmani, J. and Rasiwasia, N : “Sub-word Level Lip Reading With Visual Attention” (2021)
[9] Zhang, H., Shen, J., Huang, X., Luo, P., Liu, Q. and Zeng, W : “A New Language Independent, Photo-realistic Talking Head Driven by Voice Only” (2022)
[10] Revéret, L., Bailly, G. and Badin : “Realistic Mouth-Synching for Speech-Driven Talking Face Using Articulatory Modelling” (2022)
[11] Wang, Y., Qian, X., Soong, F.K., Jones, J.E. and Ng, M.L : “High quality lip-sync animation for 3D photo-realistic talking head” (2021)
[12] Van Den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. and Kavukcuoglu : “WaveNet: A Generative Model for Raw Audio” (2022)
[13] Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C. and Nießner, M : “Face2face: Real-time face capture and reenactment of rgb videos” (2018) 32
[14] Lan, Y., Tran, D., Ward, D. and Harvey, R : “Audio-to-Visual Speech Conversion using Deep Neural Networks.” (2021)
[15] Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W.T. and Rubinstein, M : “Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation.” (2020)
[16] Afouras, T., Chung, J.S. and Zisserman, A : “LRS3-TED: a large-scale dataset for visual speech recognition.” (2020)
[17] Chung, J.S. and Zisserman, A : “Lip reading in the wild.” (2022)